Subgroup analysis of treatment effects for misclassified biomarkers with time-to-event data

Fang Wan, Andrew C. Titman, Thomas F. Jaki

Abstract

Analysing subgroups defined by biomarkers is of increasing importance in clinical research. In many situations the biomarker is subject to misclassification error, meaning the subgroups are identified with imperfect sensitivity and specificity. In these cases, it is improper to assume the Cox proportional hazards model for the subgroup specific treatment effects for time-to-event data with respect to the true subgroups, since the survival distributions with respect to the diagnosed subgroups will not adhere to the proportional hazards assumption. This precludes the possibility of using simple adjustment procedures. Two approaches to modelling are considered; the corrected score approach of Zucker and Spiegelman (2008) and a method based on formally modelling the data as a mixture of Cox models using an EM algorithm for estimation. The methods are comparable for moderate to large sample sizes, but the EM algorithm performs better when there are 100 patients per group. An estimate of the overall population treatment effect is obtained through the interpretation of the hazard ratio as a concordance odds. The methods are illustrated on data from a renal-cell cancer trial.

1 Introduction

There is increasing acknowledgement of the existence of patient subgroups within clinical research. While some treatments work well for all patients with the same disease, it has been shown that some treatments are only effective for some subgroups of patients defined by a certain predictive biomarker [1, 2, 3, 4]. As a consequence, many clinical trials look to perform subgroup analysis to assess whether a treatment is beneficial for those patients that are biomarker positive or biomarker negative and many trial designs have been developed to account for these subgroups. Enrichment designs [5, 6] seek to identify the most promising (sub)group of patients during the study while other designs optimize the cost-efficiency of the trials via patients allocation with respect to their biomarker status (subgroup membership) [7, 8] or use a

Figure 1: A biomarker stratified design

biomarker-strategy design [9]. All of these clinical trials assume 100% accuracy of the biomarker used in defining subgroups. However, it is seldom possible to measure a biomarker with perfect diagnostic accuracy meaning the observed subgroups will be subject to misclassification error. Without taking the sensitivity and specificity of the biomarker into consideration, the resulting conclusion may be inaccurate [10, 11]. Existing methods that account for the sensitivity and specificity [10, 11] consider normal and binary endpoints only, while time-to-event data has not yet been considered.

In this paper, we propose a method to obtain point estimates and confidence intervals of the treatment effects in biomarker stratified subgroups with time-to-event data for a biomarker by treatment interaction design depicted in Figure 1 [12]: Assume the total number of patients available to be enrolled into the trial is fixed to be N. Patients are classified into two subgroups according to the observed status (positive or negative) of a specific biomarker. In each of the two subgroups, patients are randomized into either the treatment or control arm and are administered experimental treatment or placebo/active control accordingly. The primary outcome, which is the survival time subject to right censoring, of all patients enrolled are recorded for analysis.

The remainder of the article is organized as follows. In Section 2, the statistical model for misclassified biomarker subgroups is defined. Section 3 gives estimation procedures for the model parameters, measures of overall efficacy and construction of confidence intervals. Section 4 presents a simulation results to assess the performance of the estimator and confidence intervals. The method is illustrated on a data example relating to metastatic renal-cell cancer in Section 5. The article concludes with a discussion.

2 Statistical Model

Conditional on the true biomarker status, a proportional hazards model is assumed to hold. Specifically the hazard at time t for patient i is taken as

$$h_i(t; x_i, z_i) = h_0(t) \exp(\beta_1 x_i + \beta_2 z_i + \gamma x_i z_i) \tag{2.1}$$

where $h_0(t)$ is an unspecified baseline hazard function, x_i and z_i are binary indicators of treatment and true biomarker status, respectively. Note that the biomarker status is 0 for the true negative subgroup and 1 for the true positive subgroup. Further note that this model assumes that the biomarker status to be measured

without error in this model. Under this model, the hazard ratios associated with the treatment are $\exp(\beta_1)$ and $\exp(\beta_1 + \gamma)$ for patients in the biomarker negative and positive group, respectively.

When the true biomarker status cannot be observed, a diagnostic test with imperfect sensitivity and specificity has to be used. Let $v_i \in \{0, 1\}$ be a binary indicator of whether the ith patient tests positive for the biomarker. The marginal distribution of survival times among patients in each diagnosis group will then be a mixture of Cox models corresponding to the models under true biomarker positive or negative status and with the mixing proportions determined by the positive-predictive value (PPV) and negative-predictive value (NPV) of the diagnostic test.

The PPV is given by

$$p_{+|\oplus} := \frac{\pi \times \lambda_1}{\pi \times \lambda_1 + (1 - \pi)(1 - \lambda_2)} \tag{2.2}$$

and NPV by

$$p_{-|\ominus} := \frac{(1 - \pi)\lambda_2}{\pi(1 - \lambda_1) + (1 - \pi)\lambda_2} \tag{2.3}$$

where the sensitivity, λ_1, and the specificity, λ_2 are assumed to be known and the prevalence of the biomarker, π, may either be considered known or will be estimated from the data.

The survivor function for patients observed to be positive and negative are then

$$S_{\oplus}(t; x) \coloneqq S(t; x, v = 1) = p_{+|\oplus} S(t; x, z = 1) + (1 - p_{+|\oplus}) S(t; x, z = 0) \tag{2.4}$$

and

$$S_{\ominus}(t; x) \coloneqq S(t; x, v = 0) = (1 - p_{-|\ominus}) S(t; x, z = 1) + p_{-|\ominus} S(t; x, z = 0), \tag{2.5}$$

respectively, where $S(t; x, z) = \exp\{-H_0(t) \exp(\beta_1 x + \beta_2 z + \gamma x z)\}$ and $H_0(t) = \int_0^t h_0(u) du$ is the baseline cumulative hazard.

Note that unless there is either no treatment effect or the biomarker is observed without misclassification, proportional hazards will not hold with respect to the treatment x, for either $S_{\oplus}(t; x)$ or $S_{\ominus}(t; x)$. Therefore it is not possible to fit a Cox model to the observed data and perform some simple correction to adjust for misclassification error.

3 Estimation

3.1 Corrected score estimation approach

The issue of misclassification in biomarker subgroup survival analysis can be considered a special case of measurement error in covariates, for which there is a rich previous literature [13, 14, 15, 16]. While much

of the literature concentrates primarily on continuous covariates, Zucker and Spiegelman [17] proposed an estimating equations based approach for binary covariates subject to misclassification. The approach involves constructing consistent estimates of each of the constituent terms in the Cox partial likelihood score equation to obtain a corrected score equation. Here we present how the method proceeds specifically in the case of the model given in 2. The score equation involves terms of the form $G(x_i, z_i)$, but z_i cannot be observed directly. Therefore instead the method seeks a function $G^*(x_i, v_i)$ based on the observable data, such that

$$\mathbb{E}[G^*(x_i, v_i)|x_i, v_i] = G(x_i, z_i).$$

Specifically, let

$$\mathbf{A} = \begin{bmatrix} \lambda_2 & 1-\lambda_2 \\ 1-\lambda_1 & \lambda_1 \end{bmatrix}$$

be the 2×2 matrix with (l, m) entry corresponding to $\mathbb{P}(v_i = m - 1|z_i = l - 1)$. Then

$$G^*(x_i, v_i) = \sum_{l=0}^{1} \mathbf{B}_{v_i+1,l+1} G(x_i, l),$$

where $\mathbf{B} = \mathbf{A}^{-1}$.

Let $\boldsymbol{\theta} = (\beta_1, \beta_2, \gamma)^{'}$, then applying this approach to each of the partial likelihood score equations for the model in (2.1) leads to equations

$$U_1^*(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \left\{ x_i - \frac{e_x^*(t_i)}{e_0^*(t_i)} \right\}, \tag{3.1}$$

$$U_2^*(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \left\{ \mathbf{B}_{v_i+1,2} - \frac{e_z^*(t_i)}{e_0^*(t_i)} \right\}, \tag{3.2}$$

$$U_3^*(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \left\{ \mathbf{B}_{v_i+1,2} x_i - \frac{e_{xz}^*(t_i)}{e_0^*(t_i)} \right\} \tag{3.3}$$

$$, \tag{3.4}$$

where

$$e_x^*(t) = \frac{1}{n} \sum_{k=1}^{n} \sum_{l=0}^{1} Y_k(t) \mathbf{B}_{v_k+1,l+1} x_k \exp\{\boldsymbol{\theta}^{'} \mathbf{x}_k^*(l)\},$$

$$e_z^*(t) = \frac{1}{n} \sum_{k=1}^{n} Y_k(t) \mathbf{B}_{v_k+1,2} \exp\{\boldsymbol{\theta}^{'} \mathbf{x}_k^*(1)\},$$

$$e_{xz}^*(t) = \frac{1}{n} \sum_{k=1}^{n} Y_k(t) \mathbf{B}_{v_k+1,2} x_k \exp\{\boldsymbol{\theta}^{'} \mathbf{x}_k^*(1)\},$$

where $Y_k(t) = I(t_i \geq t)$ is the at risk indicator for patient k and $\mathbf{x}_k^*(l) = (x_k, l, l \times x_k)^{'}$, for $l = 0, 1$. The score equations are not unbiased, but are asymptotically unbiased and therefore taking $\hat{\boldsymbol{\theta}}$ as the solution to $\mathbf{U}^*(\boldsymbol{\theta}) = \mathbf{0}$.

An estimate of the cumulative baseline hazard can be obtained by

$$\hat{H}_0(t) = \sum_{i=1}^{n} \frac{\delta_i I(t_i < t)}{e_0^*(t_i, \hat{\boldsymbol{\theta}})}$$

which is analogous to the standard Breslow estimate.

Note that the estimator does not depend on the underlying prevalence of biomarker positive patients in the population, π. Therefore, if an estimate of π is required, one can take

$$\hat{\pi} = \frac{\bar{v} + \lambda_2 - 1}{\lambda_1 + \lambda_2 - 1}, \tag{3.5}$$

where $\bar{v} = \sum_i v_i/n$, which arises by maximizing

$$L^*(\pi) = \prod_i (\pi\lambda_1 + (1-\pi)(1-\lambda_2))^{v_i} (\pi(1-\lambda_1) + (1-\pi)\lambda_2)^{1-v_i}.$$

The corrected score method has the advantage of being computationally efficient. However, a known drawback is that solutions to the score equations will not necessarily exist, particularly for small sample sizes or when the misclassification rate is high.

3.2 Semi-parametric maximum likelihood approach

Alternatively, estimation of the model in (2) can proceed using a semi-parametric maximum likelihood approach. A full likelihood for the data can be constructed by making the standard assumption that the hazard function $h_0(t)$ is piecewise constant between observed event times [18]. The true biomarker status may be considered missing data, for which the observed biomarker status and the patient's survival can be considered indicators.

3.3 EM algorithm

Direct maximization of the likelihood is difficult or infeasible due to the large number of nuisance parameters associated with the increments of the baseline hazard. Instead, taking a similar approach to various previous authors [19, 20, 21, 22], an Expectation-Maximization (EM) algorithm is used. The true biomarker status is treated as missing data, such that the 'M'-step of the algorithm involves fitting a weighted Cox model, where each patient has two sets of data corresponding to being truly biomarker positive or biomarker negative. The

weights correspond to the conditional probability of being truly biomarker positive (or negative) given the current estimates of the parameters and the observed data (follow-up time, event indicator and diagnostic test result).

Let $t_{(j)}$ denote the jth ordered uncensored event time and $t_{(0)} = 0$, then

$$h_0(t) = h_j, \quad \text{for} \quad t_{(j-1)} < t \leq t_{(j)}.$$

The full information log-likelihood is then

$$\log L_F(\boldsymbol{\theta}, \mathbf{h}, \pi) = \sum_i \delta_i(\log h_0(t_i) + \{(\beta_1 + \gamma z_i)x_i + \beta_2 z_i\}) - H_0(t_i)\exp\{(\beta_1 + \gamma z_i)x_i + \beta_2 z_i\}$$
$$+ z_i \log \pi + (1 - z_i)\log(1 - \pi),$$

where $H_0(t)$ denotes the cumulative baseline hazard. The likelihood contributions for the ith observed subject given observed positive and negative subgroup status is

$$L_{+i} = [h_0(t_i)\exp\{(\beta_1 + \gamma)x_i + \beta_2\}]^{\delta_i} \exp[-H_0(t_i)\exp\{(\beta_1 + \gamma)x_i + \beta_2\}]$$

and

$$L_{-i} = [h_0(t_i)\exp\{\beta_1 x_i\}]^{\delta_i} \exp[-H_0(t_i)\exp\{\beta_1 x_i\}],$$

respectively.

The expectation step of the EM algorithm involves calculating $E(z_i|t_i, \delta_i, x_i, v_i, \boldsymbol{\theta}^{(l)}, \mathbf{h}^{(l)}, \pi^{(l)})$. This produces conditional weights of the form

$$\begin{aligned} w_i &:= P(z_i = 1|t_i, \delta_i, x_i, v_i, \boldsymbol{\theta}^{(l)}, \mathbf{h}^{(l)}, \pi^{(l)}) \\ &= \left(\frac{p_{+|\oplus} L_{+i}}{p_{+|\oplus} L_{+i} + (1 - p_{+|\oplus})L_{-i}}\right)^{v_i} \left(\frac{(1 - p_{-|\ominus})L_{+i}}{(1 - p_{-|\ominus})L_{+i} + p_{-|\ominus} L_{-i}}\right)^{1 - v_i}. \end{aligned}$$

Note that the weights depend on both $\boldsymbol{\theta}$, H_0, as well as π via the PPV and NPV. The expected conditional likelihood is given by

$$Q(\boldsymbol{\theta}, \mathbf{h}, \pi|\boldsymbol{\theta}^{(l)}, \mathbf{h}^{(l)}, \pi^{(l)}) = \sum_i \delta_i[\log h_0(t_i) + \{(\beta_1 + \gamma w_i)x_i + \beta_2 w_i\}] - H_0(t_i) w_i \exp\{(\beta_1 + \gamma)x_i + \beta_2\}$$
$$- H_0(t_i)(1 - w_i)\exp\{\beta_1 x_i\} + w_i \log \pi + (1 - w_i)\log(1 - \pi). \tag{3.6}$$

If the simplifying assumption that censored patients' follow-up time is taken to be directly after the previous uncensored failure time, then for a given $\boldsymbol{\theta}$, (3.6) is maximized with respect to $\mathbf{h}$ by

$$h_j = \left(\{t_{(j)} - t_{(j-1)}\} \sum_{k \in \mathcal{R}_j} w_k \exp\{(\beta_1 + \gamma)x_k + \beta_2\} + (1 - w_k)\exp\{\beta_1 x_k\}\right)^{-1}$$

and $\mathcal{R}_j = \{i : t_i \geq t_{(j)}\}$ denotes the risk set of patients at time $t_{(j)}$. Substituting this into (3.6) leads to

$$Q_h(\boldsymbol{\theta}, \pi | \boldsymbol{\theta}^{(l)}, \pi^{(l)}) = \sum_i \delta_i [(\beta_1 x_i + \beta_2 w_i + \gamma x_i w_i) - \log\{\sum_{k \in \mathcal{R}_i} w_k \exp\{(\beta_1 + \gamma)x_k + \beta_2]\} + (1 - w_k)\exp(\beta_1 x_k)\}] + w_i \log \pi + (1 - w_i)\log(1 - \pi), \tag{3.7}$$

where $\mathcal{R}_i = \{k : t_k \geq t_i\}$. The first part of Q_h is the same as a Cox partial likelihood for a weighted sample of n biomarker positive patients weighted by $\{w_i\}$ and n biomarker negative patients weighted by $\{1 - w_i\}$. As such the M-step for $\boldsymbol{\theta}$ can be computed using standard software for Cox regression. In addition the update for π is given by $\pi^{(l+)} = n^{-1} \sum_i w_i$. Note that the values of $p_{+|\oplus}$ and $p_{-|\ominus}$ are also updated by plugging the new estimate of π into (2.2) and (2.3).

The survival data contains essentially no additional information about π beyond that in the observed values of v_i. As a consequence taking the estimator of π in (3.5) gives a value very close to the maximum profile likelihood estimate of π.

3.4 Construction of confidence intervals

Corrected score method

An estimate of the variance-covariance matrix of $\hat{\boldsymbol{\theta}} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\gamma})'$ can be found through taking

$$\hat{\mathbf{V}} = \mathbf{D}(\hat{\boldsymbol{\theta}})^{-1} \mathbf{H}(\hat{\boldsymbol{\theta}}) \mathbf{D}(\hat{\boldsymbol{\theta}})^{-1}$$

where $\mathbf{D}(\boldsymbol{\theta})$ is the 3×3 matrix of derivatives of $\mathbf{U}^*(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and $\mathbf{H}(\boldsymbol{\theta})$ is an empirical estimate of the covariance matrix of $n^{1/2}\mathbf{U}^*(\boldsymbol{\beta})$. Full details are given in [17].

Given $\hat{\mathbf{V}}$, normal Wald confidence intervals can be constructed for individual parameters.

An estimate of the standard error of $\hat{\pi}$ is

$$SE(\hat{\pi}) = \frac{\sqrt{\bar{v}(1 - \bar{v})}}{\sqrt{n}(\lambda_1 + \lambda_2 - 1)},$$

and $\hat{\pi}$ is asymptotically independent of $\hat{\boldsymbol{\theta}}$.

EM algorithm approach

A convenient approach to constructing asymptotic confidence intervals for individual parameters when estimation is via the EM algorithm is based upon the profile likelihood ratio, which continues to have standard χ^2 asymptotics even in the presence of a potentially infinitely dimensional nuisance parameter [26].

The observed (or marginal) likelihood for the data is given by

$$L(\boldsymbol{\theta}, \hat{H}_0(t), \pi) = \prod_i \{P(t_i, \delta_i | x_i, v_i, \boldsymbol{\theta}, \hat{H}_0(t), \pi) P(v_i | \pi)\},$$

which may be expanded as

$$\begin{aligned} L(\boldsymbol{\theta}, \hat{H}_0(t), \pi) = \prod_i \{\pi \times \lambda_1 L_{+i} + (1-\pi) \times (1-\lambda_2) L_{-i}\}^{v_i} \times \\ \{\pi \times (1-\lambda_1) L_{+i} + (1-\pi) \times \lambda_2 L_{-i}\}^{1-v_i}. \end{aligned} \tag{3.8}$$

To obtain a confidence interval for the interaction parameter γ, for instance, we use the fact that

$$\Lambda(\hat{\gamma}, \gamma_0) = 2 \log \frac{L(\hat{\boldsymbol{\theta}}, \hat{H}_0(t))}{L(\hat{\beta}_1, \hat{\beta}_2, \gamma_0, \hat{H}_0(t))} \xrightarrow{d} \chi^2_1$$

and hence take $\{\gamma : \Lambda(\hat{\gamma}, \gamma) \leq \chi^2_1(1-\alpha)\}$ as a $(1-\alpha) \times 100\%$ confidence interval for γ. It is straightforward to find the maximum profile likelihood estimates by using a modified EM algorithm where at each M-step the fixed parameter, e.g. γ, is treated as a fixed offset term in the weighted Cox model.

For the subgroup analysis it is also desirable to construct a simultaneous confidence interval for the estimated treatment effect in the biomarker positive and negative groups in order to control the familywise type I error rate. In the parametrization used in (2.1) this corresponds to simultaneous confidence intervals for $(\beta_1 + \gamma)$ and β_1. In this case, the method proceeds by obtaining an estimate of the Hessian of the observed profile likelihood with respect to (β_1, γ). The method of [27] is used to approximate the profile likelihood information. This approach has also been used in other contexts where estimation requires an EM algorithm [28]. The profile likelihood information is approximated by computing the profile likelihood at values about $(\hat{\beta}_1, \hat{\gamma})$, perturbed by a suitably small value h to provide a 'finite-differences' type approximation. Specifically,

$$I_{\beta_1\beta_1} \approx -\frac{l_p(\hat{\beta}_1 + 2h, \hat{\gamma}) - 2l_p(\hat{\beta}_1 + h, \hat{\gamma}) + l_p(\hat{\beta}_1, \hat{\gamma})}{h^2},$$

$$I_{\beta_1\gamma} \approx -\frac{l_p(\hat{\beta}_1 + h, \hat{\gamma} + h) - l_p(\hat{\beta}_1, \hat{\gamma} + h) - l_p(\hat{\beta}_1 + h, \hat{\gamma}) + l_p(\hat{\beta}_1, \hat{\gamma})}{h^2}$$

and

$$I_{\gamma\gamma} \approx -\frac{l_p(\hat{\beta}_1, \hat{\gamma} + 2h) - 2l_p(\hat{\beta}_1, \hat{\gamma} + h) + l_p(\hat{\beta}_1, \hat{\gamma})}{h^2}.$$

The value of h is primarily chosen to ensure that numerical stability in the converged values of the EM algorithm do not affect the estimate. Theoretically, the value of h should decrease with increasing sample size, but taking $h = 0.01$ worked adequately in the examples considered in this paper and the results were not particularly sensitive to the choice of h.

Simultaneous confidence intervals

Using either the corrected score approach or profile likelihood, an estimate of $\boldsymbol{\Psi}$, the variance covariance matrix of $(\hat{\beta}, \hat{\gamma})$ can be found. For the corrected score approach this just involves taking the relevant components of $\hat{\mathbf{V}}$, while for the profile likelihood method it involves inverting the profile information matrix with respect to (β_1, γ).

An estimate of $\boldsymbol{\Sigma}$, the variance-covariance matrix of $(\hat{\beta}_1 + \hat{\gamma}, \hat{\beta}_1)$, can then be obtained by taking

$$\hat{\Sigma} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right) \boldsymbol{\Psi} \left(\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right).$$

Simultaneous confidence intervals are then constructed of the form

$$(\hat{\beta}_1 + \hat{\gamma}) \pm \xi_\alpha \sigma_+ \quad \text{and} \quad \hat{\beta}_1 \pm \xi_\alpha \sigma_-$$

where ξ_α is the scaling factor chosen such that, for a bivariate normal random variable, $\mathbf{X}$, with unit variances and correlation ρ, $P(|X_1| \leq \xi_\alpha \cap |X_2| \leq \xi_\alpha) = 1 - \alpha$. This value can be found straightforwardly using the `qmvnorm` function in the mvtnorm package in **R** [29, 30].

3.5 Missing biomarker status

In some trials, only a subset of patients may have had their biomarker status measured. If it can be assumed that the missing diagnostic tests of biomarker status are missing at random, then the survivor function for such patients, $S_\odot(t; x)$, is given by:

$$S_\odot(t; x) = \pi S(t; x, z = 1) + (1 - \pi) S(t; x, z = 0).$$

Such patients can easily be accommodated within the EM algorithm proposed in Section 3.3 by a simple modification of the conditional weights for such patients. Specifically, the weight for a patient i with missing diagnostic test is taken as

$$w_i = \frac{\pi L_{+i}}{\pi L_{+i} + (1 - \pi) L_{-i}}.$$

Similarly, the marginal likelihood contribution of these patients, regardless of whether π is taken as known or to be estimated is simply given by

$$L_i(\boldsymbol{\theta}, \hat{H}_0(t), \pi) = \pi L_{+i} + (1 - \pi) L_{-i}.$$

Since the corrected score approach does not directly use the prevalence, there does not seem an obvious way to incorporate the information on survival for those with missing biomarker status into the corrected scores.

If the diagnostic test results are missing completely at random then consistent estimates can be obtained by deleting data from patients without a diagnostic test result. However, this will lead to a loss in efficiency particularly with respect to estimating the overall treatment effect.

3.6 Measures of overall efficacy

The use of hazard ratios for subgroup analysis of time-to-event data has been criticized due to the absence of a constant hazard ratio in a mixture population and as a consequence, other methods based on median survival and parametric modelling have been proposed to obtain 'subgroup mixable' estimates [24].

However, the hazard ratio between two groups in a proportional hazards model can also be expressed as the concordance odds [25]. Specifically, if T_0 and T_1 are the survival times of two randomly chosen individuals from groups 0 and 1 and the hazard ratio of group 1 compared to group 0 is ψ, then $\frac{P(T_0>T_1)}{1-P(T_0>T_1)} = \psi$, or equivalently

$$P(T_0 > T_1) = \frac{\psi}{1+\psi}. \tag{3.9}$$

The concordance odds has a clear clinical meaning and has the advantage that an estimate of the overall concordance odds in a subgroup model can be found as a function of just the individual subgroup concordance odds and the prevalence. It also has the advantages of not requiring either fully parametric estimation procedures, which may be less robust, or fully non-parametric procedures which will be less efficient.

Let $T_i, i = 0, 1$ represent the survival time of a random subject in treatment arm i and $G_i \in \{0, 1\}, i = 0, 1$ represent the subgroup membership with $P(G_i = 1) = \pi$, then

$$\begin{aligned} P(T_0 > T_1) = {} & \pi^2 P(T_0 > T_1 | G_0 = G_1 = 1) + (1-\pi)^2 P(T_0 > T_1 | G_0 = G_1 = 0) \\ & + \pi(1-\pi) P(T_0 > T_1 | G_0 = 0, G_1 = 1) + \pi(1-\pi) P(T_0 > T_1 | G_0 = 1, G_1 = 0), \end{aligned}$$

hence

$$P(T_0 > T_1) = \pi^2 P(T_{01} > T_{11}) + (1-\pi)^2 P(T_{00} > T_{10}) + \pi(1-\pi) P(T_{00} > T_{11}) + \pi(1-\pi) P(T_{01} > T_{10})$$

where T_{ij} is the survival time for a subject in treatment arm i and subgroup j. Each of the probabilities on the right-hand side of the equation can be expressed in terms of the parameters β_1, β_2, γ of the model in (2.1). Following (3.9), we have

$$P(T_0 > T_1) = \pi^2 \text{expit}(\beta_1+\gamma) + (1-\pi)^2 \text{expit}(\beta_1) + \pi(1-\pi)\text{expit}(\beta_1+\beta_2+\gamma) + \pi(1-\pi)\text{expit}(\beta_1-\beta_2), \tag{3.10}$$

where $\text{expit}(x) = (1 + \exp(-x))^{-1}$ is the inverse logit function. An estimate of the overall effect of a treatment, expressed as concordance odds, is then given by $\frac{\hat{P}(T_0>T_1)}{1-\hat{P}(T_0>T_1)}$ where $\hat{P}(T_0 > T_1)$ is obtained by plugging the estimates of β_1, β_2, γ and π into (3.10).

A disadvantage of the concordance odds as a measure of treatment efficacy is that there is no guarantee that the overall efficacy measure lies in the interval between the two subgroup efficacy values. In fact, when $\gamma = 0$ but $\beta_2 \neq 0$ the overall concordance odds will be closer to 1 than $\exp(\beta_1)$. Nevertheless it is virtually impossible, in practice, for a contradictory result to occur, i.e. for there to be statistically significant benefits for both subgroups but a non-significant overall effect. Such a situation could only occur if $|\hat{\beta}_2|$ is large and $SE(\hat{\beta}_2)$ is large compared to both $SE(\hat{\beta}_1)$ and $SE(\hat{\beta}_1 + \hat{\gamma})$. Moreover it is impossible for the overall effect to be of a different sign to the two subgroup effects.

If desired, the procedures for constructing simultaneous confidence intervals in Section 3.4 can be extended to provide simultaneous confidence intervals for the two subgroup effects and the overall log concordance odds by computing the variance-covariance matrix of $(\hat{\beta}_1 + \hat{\gamma}, \hat{\beta}_1, \hat{\beta}^*)$ using the delta method, where $\hat{\beta}^* = \log\left\{\frac{\hat{P}(T_0>T_1)}{1-\hat{P}(T_0>T_1)}\right\}$. Obtaining an analytical form for the first derivatives of the transformation can be cumbersome, but a numerical approximation for the first derivatives can be used instead.

4 Simulations

To investigate the finite sample properties of the proposed semi-parametric likelihood estimator and compare with the corrected score estimator, data sets of varying sizes and levels of biomarker subgroup diagnostic accuracy are simulated. The underlying survival hazards are assumed to follow the model in (2.1), with a decreasing Weibull baseline hazard assumed with rate parameter 0.1 and shape parameter 0.8 such that $h_0(t) = 0.8 \times 0.1^{0.8} t^{-0.2}$.

Three scenarios are considered for the treatment effects. In the first, $\beta_1 = -0.5, \beta_2 = 0.1$ and $\gamma = 0.3$, meaning the treatment is beneficial for both biomarker groups, but the effect is smaller for those who are biomarker positive, corresponding to hazard ratios (HR) of 0.61 and 0.82. In the second, $\beta_1 = 0.1, \beta_2 = 0.1$ and $\gamma = -0.7$ corresponding to a stronger interaction effect where the treatment is beneficial for the biomarker positive group but slightly harmful for the negative group (HRs of 0.55 and 1.11). Finally the third scenario, $\beta_1 = 0, \beta_2 = 0.1$ and $\gamma = 0$, corresponds to a situation where the treatment has no effect in either biomarker group.

Censoring is assumed to be independent and uniform distributed between 5 and 25, $\mathrm{U}(5, 25)$, which results in an overall censoring rate of around 25%. The prevalence of a true positive biomarker status, π, is taken to be 0.3 and treated as unknown in the estimation procedure. The effect of assuming rather than estimating the prevalence is negligible in the simulation (where the prevalence given is accurate). However, if the assumed

prevalence value were far from its true value, this would lead to bias. The sensitivity and specificity of the diagnostic test is assumed known in all cases, but varied and the sample size per randomization group is varied across the simulation scenarios. The results using the EM algorithm approach and using the corrected score (SC) approach of Section 3.1 for 5000 replications of each scenario are presented in Tables 1, 2 and 3. Using the EM approach, the parameter estimates have reasonably low levels of bias for all scenarios considered. As would be expected, the standard deviation of the estimates increases as the diagnostic accuracy decreases. In the scenarios considered, since the prevalence is lower than 0.5, imperfect specificity has a greater impact than imperfect sensitivity. The standard deviation of the estimate of γ is around 75% higher when the sensitivity and specificity are both 0.8 compared to the case of perfect diagnostic accuracy. In the first scenario where the interaction effect is relatively modest, the power to detect the interaction term is low in all scenarios, but substantially lower when there is diagnostic error. For instance when the number in each randomization group is 500, the power reduces from 0.43, with perfect diagnostic accuracy, to 0.17 with sensitivity and specificity both 0.8. A similar pattern is observed in the second scenario, where the interaction effect is stronger. In the scenario with no interaction effect the empirical Type I error of a test of interaction is close to 5% for all configurations, with a slight tendency to be anti-conservative in the smaller sample size and when misclassification rates are higher.

For sample sizes of 500 per randomization group the estimates based on the corrected score approach are comparable with those using the EM algorithm; however, for $N = 100$, the biases are mostly higher and similarly the SDs are slightly higher. Most significantly, the correct score method fails to produce an estimate in some cases, with the frequency of non-convergence higher for scenarios with lower sensitivity and specificity and up to 4.8% of the time.

The simultaneous confidence intervals for $(\beta_1, \beta_1 + \gamma)$ have close to the nominal 95% level in all cases, with a tendency to be slightly conservative. The apparently better coverage of simultaneous confidence intervals for the corrected score method when $N = 100$ is likely due to only considering samples where an estimate was obtained.

5 Example: Pazonpanib for renal-cell cancer

As an illustrative example of the impact of accounting for misclassification of biomarkers in a survival study, data from a Phase III trial of patients with metastatic renal-cell cancer are analyzed. The trial involved 343 patients, 225 of whom were randomized to treatment with Pazopanib, with the remaining 118 on placebo.

Table 1: Bias and Standard Deviation (SD) of parameter estimates, empirical coverage of simultaneous (Simult) nominal 95% confidence intervals of $(\beta_1, \beta_1 + \gamma)$ and the empirical power of likelihood ratio test of interaction term in the mild interaction scenario $(\beta_1, \beta_2, \gamma) = (-0.5, 0.1, 0.3)$.

			Bias $\times 10^2$			SD			Coverage	Power	
N	(Sens,Spec)	Method	β_1	β_2	γ	β_1	β_2	γ	Simult	$\gamma \neq 0$	No est
100	(1,1)	Cox	0.0191	0.7232	-0.6650	0.2153	0.2634	0.3847	0.9472	0.1226	0.00
100	(1,0.8)	EM	-0.4598	0.9980	-2.0488	0.2416	0.3514	0.5154	0.9614	0.0902	0.00
		CS	-0.3064	1.7358	-0.3952	0.2452	0.3846	0.5701	0.9458	0.0951	0.78
100	(0.8,1)	EM	-0.6302	1.1681	-1.1618	0.2299	0.3017	0.4473	0.9512	0.1128	0.00
		CS	-1.0677	0.4218	0.3820	0.2315	0.3104	0.4561	0.9462	0.1050	0.00
100	(0.9,0.9)	EM	-0.6628	1.1261	-1.2436	0.2383	0.3366	0.4960	0.9622	0.0912	0.00
		CS	-0.6839	0.5528	0.7765	0.2396	0.3615	0.5276	0.9457	0.1004	0.24
100	(0.8,0.8)	EM	-0.4748	0.0643	-1.7780	0.2685	0.4767	0.6887	0.9596	0.0814	0.00
		CS	-1.4529	1.2653	0.7808	0.2902	0.5524	0.8193	0.9412	0.0754	4.80
500	(1,1)	Cox	0.1274	0.2206	-0.2761	0.0965	0.1157	0.1670	0.9506	0.4286	0.00
500	(1,0.8)	EM	-0.1734	-0.1977	-0.0315	0.1077	0.1583	0.2280	0.9546	0.2490	0.00
		CS	-0.0209	0.2658	-0.2750	0.1073	0.1572	0.2340	0.9448	0.2756	0.00
500	(0.8,1)	EM	-0.3087	-0.1782	0.4508	0.1012	0.1333	0.1948	0.9522	0.3514	0.00
		CS	-0.2403	-0.0325	0.4227	0.1010	0.1304	0.1965	0.9520	0.3470	0.00
500	(0.9,0.9)	EM	-0.2142	-0.3322	0.1361	0.1052	0.1490	0.2172	0.9520	0.2834	0.00
		CS	-0.2195	-0.0461	0.5080	0.1054	0.1490	0.2214	0.9442	0.2990	0.00
500	(0.8,0.8)	EM	-0.1233	-0.4452	-0.3492	0.1218	0.2051	0.2949	0.9578	0.1678	0.00
		CS	0.0210	0.3294	-0.3753	0.1232	0.2097	0.3086	0.9448	0.1888	0.00

In addition, patients were classified by level of interleukin 6 (IL-6) into 'low' or 'high' groups. Interest lies in determining whether Pazonpanib is an effective treatment for either or both groups of patient. In the original analysis by [31], it was assumed that the assay used to determine the level of IL-6 had 100% diagnostic sensitivity and specificity.

Here, the data are re-analysed considering the possibility of misclassification of IL-6 status. The individual level data were reconstructed from the Kaplan-Meier estimates provided in [31] using the method of [32]. Following [10], it is assumed that the assay has 95% sensitivity and 90% specificity to distinguish high IL-6 from low.

Table 2: Bias and Standard Deviation (SD) of parameter estimates, empirical coverage of simultaneous (Simult) nominal 95% confidence intervals of $(\beta_1, \beta_1 + \gamma)$ and the empirical power of likelihood ratio test of interaction term in the strong interaction scenario $(\beta_1, \beta_2, \gamma) = (0.1, 0.1, -0.7)$.

			Bias $\times 10^2$			SD			Coverage	Power	
N	(Sens,Spec)	Method	β_1	β_2	γ	β_1	β_2	γ	Simult	$\gamma \neq 0$	No est
100	(1,1)	Cox	0.1509	0.1085	-0.5872	0.2034	0.2581	0.3966	0.9468	0.4562	0.00
100	(1,0.8)	EM	-0.6065	-0.5878	1.0292	0.2210	0.3430	0.5144	0.9570	0.2968	0.00
		CS	0.3274	0.8807	-5.2078	0.2303	0.3856	0.6366	0.9521	0.2250	0.18
100	(0.8,1)	EM	0.0270	1.2441	-1.6333	0.2132	0.3004	0.4587	0.9516	0.3682	0.00
		CS	0.5297	0.9895	-2.4508	0.2128	0.3117	0.4644	0.9509	0.3607	0.20
100	(0.9,0.9)	EM	-0.4779	-0.8028	-0.1945	0.2241	0.3650	0.5453	0.9578	0.2812	0.00
		CS	0.0544	1.1910	-3.5591	0.2238	0.3572	0.5310	0.9542	0.2687	0.04
100	(0.8,0.8)	EM	-1.5560	0.6694	-1.3342	0.2530	0.4714	0.8217	0.9502	0.2004	0.00
		CS	0.6902	2.0582	-9.9004	0.2709	0.6015	0.9489	0.955	0.1140	3.66
500	(1,1)	Cox	0.1167	0.1671	-0.5071	0.0900	0.1148	0.1705	0.9522	0.9878	0.00
500	(1,0.8)	EM	0.1796	0.3993	-0.3036	0.0967	0.1480	0.2173	0.9566	0.9042	0.00
		CS	0.0853	-0.1995	-0.4834	0.0980	0.1565	0.2339	0.9552	0.8514	0.00
500	(0.8,1)	EM	0.1194	0.1749	-0.1344	0.0926	0.1306	0.1975	0.9536	0.9452	0.00
		CS	-0.0742	-0.3706	0.0946	0.0959	0.1335	0.2033	0.9434	0.9388	0.00
500	(0.9,0.9)	EM	0.1950	0.4669	-0.4286	0.0981	0.1563	0.2282	0.9556	0.8772	0.00
		CS	0.0896	0.2334	-0.8122	0.0983	0.1509	0.2251	0.9514	0.8846	0.00
500	(0.8,0.8)	EM	-0.0108	0.6640	-0.0728	0.1126	0.2010	0.2959	0.9600	0.6752	0.00
		CS	0.1645	0.3244	-1.9795	0.1158	0.2032	0.3081	0.9528	0.6333	0.02

Table 4 compares the results of an analysis assuming no misclassification with estimates using the proposed method. It is seen that the effect of adjusting for misclassification is to increase the estimated interaction effect from -0.53 to -0.72, which also leads to the interaction being considered significant ($p = 0.036$). The corrected score method gives similar estimates to the EM approach but with slightly wider confidence intervals.

Table 5 gives the estimates and simultaneous 95% confidence intervals for the concordance odds of Pazonpanib for Low and High IL-6 patients. For both the original and misclassification analyses, the confidence interval for Low IL-6 includes 1, implying no treatment effect, whilst the confidence interval for High IL-6 is entirely

Table 3: Bias and Standard Deviation (SD) of parameter estimates, empirical coverage of simultaneous (Simult) nominal 95% confidence intervals of $(\beta_1, \beta_1 + \gamma)$ and the empirical Type I error of likelihood ratio test of interaction term in the null scenario $(\beta_1, \beta_2, \gamma) = (0, 0.1, 0)$.

			Bias $\times 10^2$			SD			Coverage	Type I err	
N	(Sens,Spec)	Method	β_1	β_2	γ	β_1	β_2	γ	Simult	$\gamma \neq 0$	No est
100	(1,1)	Cox	0.1389	0.0680	0.5389	0.2012	0.2629	0.3678	0.9508	0.0496	0.00
100	(1,0.8)	EM	-0.1232	0.5505	0.1944	0.222	0.3356	0.4798	0.9580	0.0588	0.00
		CS	-0.2658	0.9895	0.3645	0.2236	0.3780	0.5497	0.9462	0.0526	0.44
100	(0.8,1)	EM	-0.0146	0.2379	1.2609	0.2155	0.3109	0.4396	0.9492	0.0620	0.00
		CS	0.0511	-0.0303	-0.0578	0.2126	0.3092	0.4304	0.9476	0.0516	0.00
100	(0.9,0.9)	EM	-0.0436	-0.8576	1.3769	0.2257	0.3644	0.5110	0.9590	0.0588	0.00
		CS	-0.3642	1.0238	0.8097	0.2187	0.3590	0.5083	0.9508	0.0552	0.06
100	(0.8,0.8)	EM	-0.3072	0.6792	0.1749	0.2541	0.4698	0.7010	0.9608	0.0662	0.00
		CS	-0.0386	0.4697	0.0128	0.2661	0.5555	0.8362	0.9444	0.0477	3.56
500	(1,1)	Cox	0.1800	0.4851	-0.4689	0.0899	0.1171	0.1657	0.9436	0.0572	0.00
500	(1,0.8)	EM	0.2906	0.3950	-0.4087	0.0971	0.1480	0.2067	0.9528	0.0482	0.00
		CS	-0.0667	0.0301	0.1856	0.0996	0.1561	0.2192	0.9520	0.0530	0.00
500	(0.8,1)	EM	0.1894	0.1761	-0.0657	0.0928	0.1307	0.1848	0.9526	0.0508	0.00
		CS	0.0961	0.2181	-0.3371	0.0941	0.1333	0.1866	0.9502	0.0434	0.00
500	(0.9,0.9)	EM	0.3246	0.4625	-0.5343	0.0994	0.1563	0.2196	0.9556	0.0472	0.00
		CS	0.0532	-0.0989	0.0093	0.0985	0.1488	0.2100	0.9474	0.0544	0.00
500	(0.8,0.8)	EM	0.3481	0.6671	-0.630	0.1128	0.2011	0.2832	0.9622	0.0476	0.00
		CS	0.3132	0.8912	-0.7429	0.1141	0.2072	0.2932	0.9424	0.0604	0.00

below 1, indicating a treatment benefit.

6 Conclusion and Discussion

In this paper, we investigate subgroup analysis for time-to-event responses in biomarker stratified subgroups with misclassificated biomarkers using a proportional hazards model. Two approaches to point estimation and the construction of (simultaneous) confidence intervals for the treatment effects in biomarker subgroups in the form of the log-hazard ratio are provided. It is shown by simulation that the bias of the estimators and the coverage probabilities of the simultaneous confidence intervals are acceptable for all considered simulation

Table 4: Comparison of estimates from original Cox model analysis assuming no biomarker misclassification and model assuming 95% sensitivity and 90% specificity to detect High IL-6 via EM algorithm method and Corrected score (CS) method

	Original analysis			EM method			CS method		
Parameter	Est	95% CI	p	Est	95% CI	p	Est	95% CI	p
Pazonpanib (β_1)	-0.15	(-0.58, 0.27)	0.48	-0.12	(-0.57, 0.33)	0.58	-0.14	(-0.60, 0.32)	0.56
High IL-6 (β_2)	1.18	(0.73, 1.62)	< 0.001	1.50	(0.96, 2.10)	< 0.001	1.46	(0.94, 1.97)	< 0.001
Interaction (γ)	-0.53	(-1.08, 0.03)	0.06	-0.72	(-1.40, -0.05)	0.04	-0.68	(-1.32, -0.03)	0.04
Prevalence (π)	-	-	-	0.47	(0.41, 0.53)	-	-	-	-

Table 5: Simultaneous 95% confidence intervals for effect of Pazonpanib on overall survival for Low IL-6 patients, High IL-6 patients and all patients; CO=Concordance odds

	Original analysis		EM method		CS method	
Group	CO	95% CI	CO	95% CI	CO	95% CI
Low IL-6	0.86	(0.52, 1.41)	0.88	(0.52, 1.50)	0.87	(0.51, 1.51)
High IL-6	0.51	(0.33, 0.77)	0.43	(0.25, 0.75)	0.44	(0.28, 0.70)
All	0.70	(0.51 , 0.95)	0.67	(0.50 , 0.92)	0.68	(0.50 , 0.92)

scenarios in the case of the EM algorithm approach. The corrected score method performs comparably in the case where $N = 500$ and has computational advantages. However, for $N = 100$ the corrected score method suffers from non-convergence issues and lower efficiency.

It is apparent from the simulation results that the power to detect a subgroup effect of treatment is diminished in the presence of misclassification. Further work would be to develop sample size formulas which would allow survival trials to be adequately powered to perform subgroup analysis in the presence of biomarker misclassification.

The interpretation of a hazard ratio as the concordance odds allows an overall treatment effect estimate to be computed in subgroup analyses of time-to-event data. While the focus of this paper has been cases with misclassification of the biomarker status, the use of concordance odds can also be applied in the simpler case where the biomarker status is perfectly observed.

Acknowledgments

This work is an independent research arising in part from Prof Jaki's Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. Funding for this work was also provided by the Medical Research Council (MR/M005755/1). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, or the Department of Health.

References

[1] Sacks, F.M., Pfeffer, M.A., Moye, L.A., Rouleau, J.L., Rutherford, J.D., Cole, T.G. and Braunwald, E. (1996). The effect of Pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *New England Journal of Medicine* **335**, 1001-1009.

[2] Jackson, R.D., LaCroix, A.Z., Gass, M., Wallace, R.B., Robbins, J., Lewis, C.E. and Barad, D. (2006). Calcium plus vitamin D supplementation and the risk of fractures. *New England Journal of Medicine* **354**, 669-683.

[3] Jimeno, A., Messersmith, W.A., Hirsch, F.R., Franklin, W.A. and Eckhardt, S.G. (2009). KRAS mutatioons and sensitivity to epidermal growth factor receptor inhibitors in colorectal cancer: practical application of patient selection. *Journal of Clinical Oncology* **27**, 7, 1130-1136.

[4] Peeters, M., Douillard, J.Y., Van Cutsem, E., Siena, S., Zhang, K., Williams, R. and Wiezorek, J. (2013). Mutant KRAS codon 12 and 13 alleles in patients with metastatic colorectal cancer: assessment as prognostic and predictive biomarkers of response to panitumumab. *Journal of Clinical Oncology* **31**, 6, 759-765.

[5] Freidlin, B. and Korn, E.L. (2014). Biomarker enrichment strategies: matching trial design to biomarker credentials. *Nat Rev Clin Oncol* **11**, 2, 81-90.

[6] Magnusson, B.P. and Turnbull, B.W. (2013). Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine* **32**, 16, 26952714.

[7] Wason, J.M.S., Abraham, J.E., Baird, R.D., Gournaris, I., Vallier, A.L., Brenton, J.D., Earl, H.M. and Mander, A.P. (2015). A Bayesian adaptive design for biomarker trials with linked treatments. *British Journal of Cancer* **113**, 699-705.

[8] Mehta, C., Schafer, H., Daniel, H. and Irle, S. (2014). Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Statistics in Medicine* **33**, 26, 45154531.

[9] Kunz, C., Jaki, T. and Stallard N. An alternative method to analyse the biomarker-strategy design. Submitted.

[10] Liu, C., Liu, A., Hu, J., Yuan, V. and Halabi, S. (2014). Adjusting for misclassification in a stratified biomarker clinical trial. *Statistics in Medicine* **33**, 3100-3113.

[11] Wan, F., Kunz, C. and Jaki, T. (2018). Confidence regions for treatment effects in subgroups in biomarker stratified designs. *Biometrical Journal*. DOI: 10.1002/bimj.201700303.

[12] Gosho, M., Nagashima, K. and Sato, Y. (2012). Study designs and statistical analyses for biomarker research. *Sensors* **12**, 8966-8986.

[13] Hu, C. and Lin, D. (2002). Cox regression with covariate measurement error. *Scandinavian Journal of Statistics* **29**(4), 637-655.

[14] Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69**, 331342.

[15] Hu, P., Tsiatis, A.A. and Davidian, M. (1998). Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics* **54**, 14071419.

[16] Zucker, D.M., 2005. A PseudoPartial Likelihood Method for Semiparametric Survival Regression With Covariate Errors. *Journal of the American Statistical Association* **100**(472), 1264-1277.

[17] Zucker, D.M., and Spiegelman, D. (2008). Corrected score estimation in the proportional hazards model with misclassified discrete covariates. *Statistics in Medicine* **27**, 1911-1933.

[18] Breslow, N.E. (1972). Discussion of the paper by D.R. Cox. *Journal of the Royal Statistical Society; Series B* **34**, 216-217.

[19] Zhong, M., Sen, P.K. and Cai, J. (1996). Cox regression model with mismeasured covariates or missing covariate data. *ASA Biometrics Section Proceedings*, 323-328.

[20] Martinussen, T. (1999). Cox regression with Incomplete Covariare Measurements using the EM- algorithm. *Scandinavian Journal of Statistics* **26**, 479-491.

[21] Herring, A.H. and Ibrahim, J.G. Likelihood-Based Methods for Missing Covariates in the Cox Proportional Hazards Model. *Journal of the American Statistical Association* **96**, 292-302.

[22] Wu, R.F., Zheng, M. and Yu, W. (2016). Subgroup Analysis with Time-to-Event Data Under a Logistic-Cox Mixture Model. *Scandinavian Journal of Statistics* **43**, 863878.

[23] Bilmes, J.A. (1998). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. *International Computer Science Institute; Technical Report*, TR-97-021.

[24] Ding, Y., Lin, H-M. and Hsu J.C. (2015). Subgroup mixable inference on treatment efficacy in mixture populations, with an application to time-to-event outcomes. *Statistics in Medicine* **35**, 1580-1594.

[25] Schemper, M., Wakounig, S. and Heinze, G. (2009). The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine* **28**, 24732489.

[26] Murphy, S.A. and van der Vaart, A.W. (1997). Semiparametric likelihood ratio inference. *Annals of Statistics* **25**, 1471-1509.

[27] Murphy, S.A. and van der Vaart, A.W. (1999). Observed information in semiparametric models. *Bernoulli* **5**, 381-412.

[28] Xu, C., Baines, P.D. and Wang, J.L. (2014). Standard error estimation using the EM algorithm for the joint modeling of survival and longitudinal data. *Biostatistics* **15**, 731-744.

[29] Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T. (2014). mvtnorm: Multivariate Normal and t Distributions. *R package version 0.9-9997.* `http://CRAN.R-project.org/package=mvtnorm`

[30] Genz, A. and Bretz, F. (2009). Computation of Multivariate Normal and t Probabilities. *Lecture Notes in Statistics* **195**. Springer-Verlage, Heidelberg.

[31] Tran, H.T., Liu, Y., Zurita, A.J., Lin, Y., Baker-Neblett, K.L., Martin, A.M., Figlin, R.A., Hutson, T.E., Sternberg, C.N., Amado, R.G. and Pandite, L.N. (2012). Prognostic or predictive plasma cytokines and angiogenic factors for patients treated with pazopanib for metastatic renal-cell cancer: a retrospective analysis of phase 2 and phase 3 trials. *The Lancet Oncology* **13**, 827-837.

[32] Guyot, P., Ades, A., Ouwens, M.J. and Welton, N.J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology* **12**:9.